PROBABILIDADES Y ESTADÍSTICA (C)

Práctica 6
Aula + Laboratorio

1. Los siguientes valores son mediciones del peso (en miles de toneladas) de grandes tanques de petróleo.

   229, 232, 239, 232, 259, 361, 220, 260, 231, 229, 249, 254, 257, 214, 237, 253, 274, 230, 223, 253, 195, 269, 231, 268, 189, 290, 218, 313, 220, 270, 277, 375, 222, 290, 231, 258, 227, 269, 220, 224.

   a) Construir un esquema de tallo-hoja, en el cual los tallos sean 18, 19, 20, ...

   b) Construir una tabla de frecuencias que conste de 9 intervalos de igual longitud, siendo el primero $[175, 200)$.

   d) Graficar el histograma correspondiente a la tabla hallada en (b) de manera que el área sea la frecuencia relativa.

   e) ¿Se distribuye el conjunto de datos en forma de campana o uniformemente?

2. Consideremos $x_1, \ldots, x_n$ una muestra de una población cualquiera. Sean $\bar{x}$ y $\tilde{x}$ la media y la mediana muestral, respectivamente.

   a) Si se suma una constante c a cada uno de los x_i de la muestra, obteniéndose $y_i = x_i + c$, ¿cómo se relacionan $\bar{x}$ con $\bar{y}$ y $\tilde{x}$ con $\tilde{y}$?

   b) Si cada x_i es multiplicado por una constante c, obteniéndose $y_i = cx_i$, responder a la pregunta planteada en (a).

3. Sea s_X^2 la varianza muestral correspondiente a la muestra $x_1, \ldots, x_n$. Demostrar que:

   a) $s_X^2 = \frac{1}{n-1} \sum_{i=1}^n x_i^2 - \frac{n}{n-1} \bar{x}^2$.

   b) Si $y_i = x_i + c$, con c constante, entonces $s_Y^2 = s_X^2$.

   c) Si $y_i = cx_i$, con c constante, entonces $s_Y^2 = c^2 s_X^2$.

4. Sea $x_1, \ldots, x_n$ una muestra de una población con media μ y mediana $\tilde{\mu}$.

   (a) ¿Para qué valores de c se minimiza $\sum_{i=1}^n (x_i - c)^2$?
   (Sugerencia: derivar con respecto a c).

   (b) Usando (a) decidir cuál de estas dos cantidades es más pequeña:

   $$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \qquad \text{o} \qquad \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 .$$

   (c) ¿Para qué valores de c se minimiza $\sum_{i=1}^n |x_i - c|$?
   (Sugerencia: 1: ¿Se puede usar la misma técnica que en a)? 2: Para fijar ideas, comience con una muestra de tamaño $n = 3$).

5. Dada una muestra $x_1, ..., x_n$, se define la siguiente medida de dispersión:

$$\text{MAD}\,(x_1, \ldots, x_n) = \text{med}_{1 \le i \le n}\,|x_i - \tilde{x}|$$

   a) Sea c una constante. Si definimos $y_i = x_i + c$, ¿cuál es la relación entre MAD $(x_1, \ldots, x_n)$ y MAD $(y_1, \ldots, y_n)$?

   b) Responder (a) para $y_i = cx_i$.

   c) Calcular la MAD para los datos del Ejercicio 1.

6. A partir de una muestra $x_1, ..., x_n$ se calculan la media y el desvío estándar muestrales, $\bar{x}$ y s_X respectivamente. Si definimos $y_i = (x_i - \bar{x})\,/s_X$, ¿cuánto valen $\bar{y}$ y s_Y? Interpretar este resultado.

7. Los siguientes valores representan las ganancias, expresadas como porcentajes de ventas, de 22 firmas:

5.3	4.0	12.5	3.0	3.9	6.4	5.2	2.6	12.8	7.1	3.7
4.4	3.5	3.4	3.2	5.6	3.2	3.4	6.2	4.0	2.5	3.4

   a) Hallar la mediana muestral y los cuartiles inferior y superior de estos datos.

   b) Construir un box-plot e identificar los puntos extremos.

8. La siguiente tabla contiene valores de población, en cientos de miles, de las 10 ciudades más pobladas de 4 países en el año 1967.

Argentina	**EEUU**	**Holanda**	**Japón**
29.66	77.81	8.68	110.21
7.61	35.50	7.31	32.14
6.35	24.79	6.02	18.88
4.10	20.02	2.64	16.38
3.80	16.70	1.75	13.37
2.75	9.39	1.72	11.92
2.70	9.38	1.51	10.71
2.69	8.76	1.42	7.80
2.51	7.63	1.31	7.70
2.44	7.50	1.29	7.00

   Estos datos se encuentran en la página de la materia [1].

   a) Construir un box-plot para los datos de cada país e identificar los puntos extremos en cada caso.

   b) Comparar los centros de cada población, sus dispersiones y su simetría. ¿Cuál es el país más homogéneamente habitado?

9. Los datos graficados en la última página corresponden a las máximas concentraciones diarias (en partes por mil millones) de dióxido de azufre en Bayonne desde noviembre de 1969 hasta octubre de 1972, agrupadas por mes. Los box-plots se realizaron en base a los 36 grupos (meses) de 30 mediciones cada uno.

[1] `http://www.dm.uba.ar/materias/probabilidades_estadistica_C/2004/1/datos/`

a) ¿Aumenta o disminuye la concentración media de dióxido de azufre a través del tiempo?

b) ¿Cómo evolucionan los cuartiles superiores a lo largo del tiempo? Compare con (a).

c) ¿Qué observa en los meses de invierno? (Recuerde que se trata del hemisferio Norte).

d) ¿Qué ocurre con la dispersión de los datos cuando el nivel general de concentración es alto?

(Fuente: Rice, J. (1988). *Mathematical Statistics and Data Analysis*. Ed. Wadsworth and Brooks/Cole)

10. Dos métodos fueron usados para determinar la temperatura de fusión del hielo (Natrella, 1963). Los investigadores querían saber si los dos métodos diferían o no. Los datos siguientes dan el cambio en calor total (en calorías por gramo de masa) al pasar de hielo a -72°C a agua a 0°C.

    (Fuente: idem Ejercicio 9)

    Método A: 79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04, 79.97, 80.05, 80.03, 80.02, 80.00

    Método B: 80.02, 79.94, 79.98, 79.97, 79.97, 80.03, 79.95, 79.97

    Realice un box-plot para cada uno de los métodos en un mismo par de ejes cartesianos (como el gráfico del Ejercicio 9). A partir de los box-plots obtenidos, ¿qué le diría a los investigadores?

11. Realice un qq-plot para los datos del Ejercicio 1. ¿Qué conclusiones puede obtener sobre la distribución de los datos?

12. Este ejercicio es para familiarizarse con los qq-plots, para darse una idea de cómo pueden ser éstos cuando la distribución subyacente es normal y cuando no lo es.

    (a) Generar muestras de tamaño 25, 50 y 100 de una distribución normal. Construir qq–plots para cada una de ellas. Repetir varias veces para darse una idea de cómo se comportan los qq–plots cuando la distribución subyacente es normal.

    (b) Repetir a) para una $\Gamma(5, \frac{1}{2})$.

    (c) Repetir a) para $Y = \dfrac{Z}{U}$ donde $Z \sim N(0,1)$ y $U \sim \mathcal{U}(0,1)$ independientes.

    (d) Repetir a) para una distribución uniforme.

    (e) Repetir a) para una distribución exponencial.

    (f) ¿Puede distinguir entre la distribución normal del ítem a) y las siguientes distribuciones que no son normales?

13. Con el fin de determinar cuál sería un mejor suplemento dietario, se realizó una comparación de la retención de dos formas de hierro: Fe2+ y Fe3+. Los investigadores dividieron aleatoriamente a 36 ratas en dos grupos de igual número. A un grupo se le suministró en forma oral una concentración de 1.2 millimolar de Fe2+ y al otro grupo

se le suministró la misma concentración de Fe3+. Al cabo de cierto tiempo se realizó un conteo en cada rata para determinar el porcentaje de hierro retenido.

El archivo de datos se encuentran en la página de la materia.

(a) Realice los boxplots y los qq-plots de los porcentajes obtenidos para cada grupo. En base a estos gráficos, ¿le parece razonable suponer que cada uno de los conjuntos de datos provienen de una distribución normal? ¿Por qué?

(b) En una segunda etapa, los investigadores transformaron los datos aplicando la función logaritmo natural (ln) a cada una de las observaciones. Repita el análisis anterior con los datos transformados. En base a la información obtenida, ¿le parece razonable suponer que cada uno de los conjuntos de datos transformados se distribuyen según una distribución normal? ¿Por qué?

14. El archivo `CPU.txt`, que se encuentra en la página de la materia contiene tiempos de CPU (en segundos) correspondientes a 1000 tiempos de trabajos enviados por una consultora. Para este conjunto de datos:

(a) Calcular la media muestral, la mediana muestral y la media α-podada con $\alpha = 0.10$ (10%).

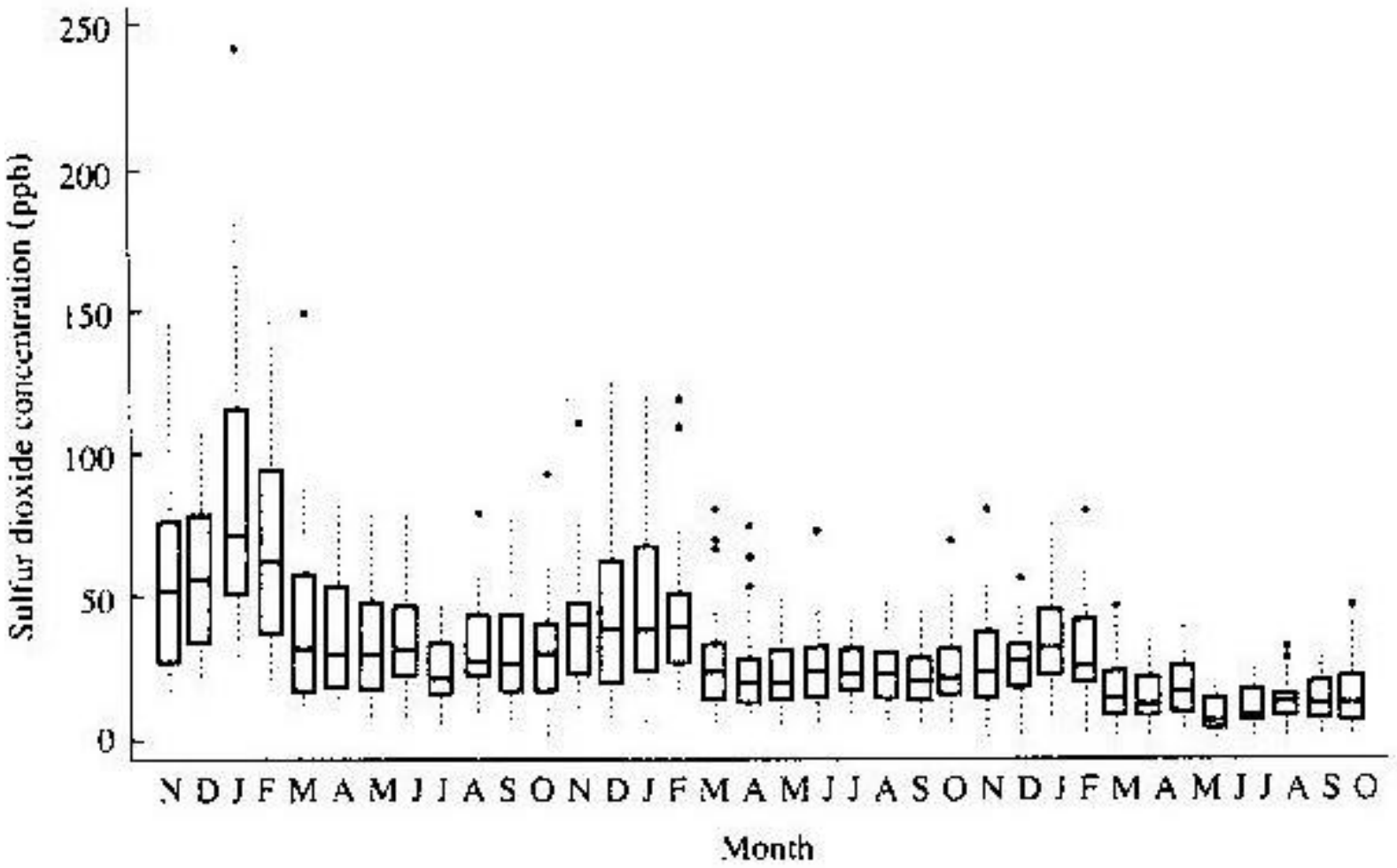
(b) Calcular el desvío estándar muestral, la distancia intercuartil y la MAD

(c) Realizar un histograma y un boxplot. ¿Cuáles son las características más sobresalientes? ¿Hay outliers?

(d) ¿Qué medida de posición cree que es más apropiada para estos datos?

(e) ¿Qué distribución cree que tienen estos datos?

(f) ¿Cómo haría para verificar si su conjetura es razonable? (SUGERENCIA: Deje volar su imaginación)


250
200
150
100
50
0
Sulfur dioxide concentration (ppb)
N D J F M A M J J A S O N D J F M A M J J A S O N D J F M A M J J A S O
Month